Détection et Reconnaissance des Gestes Emblématiques

R. Chellali, I. Renna, E. Bernier, Cyrille Achard

Détection et Reconnaissance des Gestes Emblématiques

R. Chellali[1] I. Renna[2] E. Bernier[1,3]

[1]PAVIS dept- Italian Institute of Technology
[2]ISIR- Université P&M Curie
[3]UT de Compiègne

ryad.chellali@iit.it, ilaria.renna@upmc.fr, emmanuel.bernier@iit.it

Résumé

Dans cette contribution, nous présentons un système de reconnaissance en ligne de gestes emblématiques et son utilisation pour le contrôle d'un robot mobile. Ce système a pour objectif de réduire la phase d'apprentissage d'un utilisateur humain en optimisant la reconnaissance des primitives gestuelles. En d'autres termes, nous avons traité du problème de la généralisation : faire l'apprentissage sur une base réduite de personnes et utiliser cette connaissance pour reconnaître n'importe quelle personne, indépendamment de sa morphologie, de son âge, de son sexe et de son positionnement par rapport au capteur. L'indépendance à la variabilité des personnes obtenue ici est utilisée pour réduire l'effort de mise au point d'un système de reconnaissance. Elle permet aussi d'utiliser la même approche de codage afin de générer des mouvements synthétiques plausibles qu'un agent artificiel peut exhiber dans ses interactions avec un humain. Le système obtenu comporte quatre sous-systèmes : un premier qui permet de détecter la personne et d'extraire les mouvements de la partie supérieure de cette personne. Un second, permet d'isoler les mouvements, un troisième permet de reconnaître un des mouvements appris a priori. Enfin le dernier module, traduit les mouvements reconnus en termes de contrôle d'un robot mobile à roues. Les résultats obtenus sont encourageants : ils sont conformes à l'état de l'art avec des réductions appréciables de la base d'apprentissage. Ces résultats ainsi que l'utilisation du système dans des scénarios de contrôle d'un robot sont détaillés. Nous finissons notre contribution par dresser les perspectives de l'utilisation de notre approche pour l'animation d'un agent virtuel

Mots Clef
Reconnaissance de gestes, interaction robotique, variabilité des utilisateurs, animation d'agents virtuels.

Abstract

In this paper we present a system we developed toward online emblematic gestures recognition and its use in controlling a mobile robot. The main contribution of this work is dealing with generalization: the system is able to handle a large number of performers even if the learning phase is concerned with a small number of subjects, regardless to performers' morphology characteristics, their age, their gender and their pose with regard to the sensing system. The developed approach can be seen as a classical tool allowing emblematic gestures recognition as well as a straightforward technique to generate plausible movements that artificial agents may display to human recipients. The system is composed of four sub-systems. The first one detects and extracts the upper body movements. The second sub-system segments the upper body movements. The third one allows the recognition of a set of learnt gestures' repertoire. The last module translates in terms of mobile robotics controls. We give some recognition results as well as the use of the system in handling some scenarios involving the control of a mobile robot. We finish our contribution by discussing the usage of our system in virtual agents animation context.

Keywords
Gestures recognition, interactive robotics, users' variability, plausible virtual agents animation.

Introduction

La reconnaissance de gestes est un thème traité par une large communauté allant de la sociologie à la robotique. L'intérêt pour ce thème vient surtout de son importance dans les relations et les interactions entre les humains ou entre humains et ordinateurs ou encore entre humain et robots. Des évidences concernant les relations étroites entre les actions sensorimotrices et leurs représentations dans le cerveau poussent à reconsidérer les gestes non seulement comme des supports de communication mais aussi comme un moyen de réaliser des stimulations cognitives. En d'autres termes, un geste exécuté par un agent (humain ou artificiel) peut être vu par un receveur humain de deux façons : consciemment comme un ordre ou une indication, et inconsciemment comme un programme neuro-moteur simulant le geste observé et in fine, un stimulant pour le système moteur du receveur. Ainsi, le geste est un acte dual avec une dimension linguistique (avec sa sémantique propre) et une dimension mécaniste (avec une cinématique propre pour le système polyarticulé qu'est le corps), que Kendon [1] désigne sous

le terme « morpho-kinetics ». Il associe sous ce terme les notions de phrases et de cinématique. Cette association est largement utilisée dans les travaux portant sur la compréhension du comportement humain. En effet et au delà du seul geste, plusieurs canaux de communications sont utilisés pour étudier la communication de l'homme vers un agent artificiel: la voix, les expressions faciales, les gestes ou la combinaison de tout ou partie de ces canaux. Cependant, la variabilité interpersonnelle, l'une des difficultés majeures inhérentes à chacune des techniques précédentes rendent difficiles l'implémentation de systèmes robustes et effectifs. Sur le versant machine vers l'homme, la variabilité se traduit par la difficulté pour l'agent d'exhiber des comportements gestuels plausibles. En effet, en l'absence de 'traducteurs robustes', on se retrouve avec des apparences stéréotypés souvent issus de pré enregistrements obtenus à partir de captures de mouvements. Des travaux importants [14] abordent la codification du geste, de l'expression faciale dans la perspective agent artificiel, et une partie de notre contribution abonde dans ce sens.

Dans notre cas, nous avons traité du seul geste emblématique [10], dit aussi conventionnel (voir Kendon pour une taxonomie complète [1]). Par ce dernier on entend le geste qui s'appuie uniquement sur le mouvement d'un ou deux bras pour signifier un état, un ordre ou une désignation de lieu et ce pour une large population.

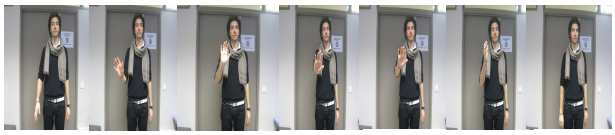


Figure 1 Un exemple d'exécution du geste STOP.

Un 'STOP' (Figure 1) par exemple sera compris par une majorité de la population comme un ordre ayant pour objectif d'arrêter l'action courante (un mouvement de marche par exemple). De même, il existe un certain nombre de gestes qui permettent de convoyer un sens ou un ordre. Dans notre travail, nous nous sommes concentrés sur le développement d'un système robuste qui permet de reconnaître et d'interpréter un certain nombre de ces gestes. Du point de vue cinématique, ces gestes peuvent être vues comme des séries temporelles décrivant les mouvements d'un système poly-articulé. Plusieurs difficultés se présentent alors dans la mise en œuvre d'un système capable de reconnaître ces séries. La première difficulté alors concerne la variabilité des séries. En effet, les mouvements sont exécutés par des personnes ayant des caractéristiques différentes et les mouvements résultants dépendent de la morphologie, de tonicité musculaire et même du sexe. Par ailleurs, ces gestes ne sont pas structurellement contraints : la désignation d'un lieu ou d'un objet (le pointage) peut être réalisée par un bras complètement tendu (alignement du bras, de l'avant bras et de la main avec l'index) vers l'objectif ou simplement par un alignement approximatif. De même, un 'BYE-BYE' peut avoir un nombre de cycles indéterminé. De ce qui précède, on peut aisément déduire que le problème de reconnaissance de gestes est double : en plus du problème de la seule reconnaissance, il est nécessaire de le traiter dans sa dimension 'variabilité'.

Dans ce travail nous n'aborderons pas à proprement parler de la synthèse gestuelle. Cependant, nous discuterons de l'intérêt de notre codage du geste dans une optique de gestes plausibles.

Dans la section suivante, nous donnons quelques éléments sur les travaux antérieurs sur le sujet traité. Nous détaillerons dans les sections 3 et 4 la méthode que nous avons mise au point pour reconnaître des gestes en ligne. Dans la section 5 nous donnons quelques exemples de l'utilisation du système pour contrôle un robot. Nous finirons par une conclusion et quelques perspectives sur les développements futurs.

Contexte et Travaux antérieurs

L'humain utilise les gestes seuls ou combinés avec d'autres modalités (voix, expression faciale, etc.) pour transmettre des concepts, des idées et des contrôles à d'autres humains. Dans la littérature, très fournie par ailleurs, des psycholinguistes, des sémioticiens, des sociologues, anthropologues, etc. le geste est considéré comme une composante essentielle pour l'établissement aussi bien des simples communications que des relations sociales complexes. On peut voir aussi dans cette littérature, toute la complexité du concept de geste et la difficulté de le cerner : 'tout le monde sait ce qu'est un geste, mais personne ne peut le définir avec précision' [2]. De fait, l'absence d'un modèle calculatoire constitue une barrière difficile. Notre ambition étant de contrôler un robot via des gestes, nous restreindrons notre analyse à une sous classe de gestes.

2.1 Les définitions du geste

Dans notre travail, nous avons abordé les seuls gestes qualifiés d'autonomes selon la terminologie de Kendon [1], où l'auteur fait la différence entre le geste autonome (ou auto-suffisant) et la gesticulation. Le premier décrit les mouvements du corps qui se suffisent à eux même en tant que porteurs d'information. Le second est plus général et concerne les mouvements qui ont besoin de compléments vocaux ou autres pour donner un sens à l'information transmise. McNeil [3] quant à lui associe le geste à la parole : il propose quatre catégories (iconique, métaphorique, déictique et le battement) qui font référence à des évènements, à des concepts abstraits, à des orientations et enfin, à des discontinuités. Cette classification ne fait pas référence à l'aspect mécaniste comme la précédente.

Cependant, elle la complète car elle supporte la dimension sémantique. A mi-chemin entre les deux, nous retrouvons

les travaux de Stokoe [12], le pionnier de la retranscription de l'ASL (American Sign Language), qui décrit le geste comme une entité ayant trois caractéristiques : la forme de la main, le lieu (géométrique) où se produit le geste et enfin sa cinématique (ou le mouvement lui même). Les travaux cités avant émanent essentiellement de psycholinguistes ou sémioticiens et concernent donc des interactions interhumaines. Dans une perspective d'interaction avec une machine, il nous a semblé important de citer le travail de Karam et Al. [4]. Ils proposent une classification plutôt dictée par les contraintes techniques et technologiques. Sans renier les travaux précurseurs dans [2] et [3], les auteurs organisent les classes en quatre catégories aussi : en premier, ils utilisent le critère 'application' ce qui donne de fait un poids important au contexte dans lequel a lieu l'interaction. En second, c'est la technologie supportant la reconnaissance des gestes incluant essentiellement les outils de mesure. En troisième, ils considèrent le type de réponse fournie par le système avec lequel il y a interaction (la fermeture de la boucle sensorimotrice). Enfin ils prennent en compte le style/forme du geste.

Pour notre problème et au vu des définitions précédentes, nous considérons en premier lieu la forme du geste comme principale caractéristique (*morpho-kinetics*). En effet, notre technologie de mesure est sans contact (l'utilisateur n'embarque aucun instrument) et permet d'extraire seulement le mouvement. Ce dernier contrôle un robot, qui en retour, fournit des réponses aux gestes par ses actions et ses déplacements.

2.2 La structure du geste, son extraction et son codage.

L'humain a la capacité de comprendre un geste emblématique exécuté par une personne qu'il n'a jamais vu, qui ne lui est pas destiné, même à travers une vidéo et sans avoir le contexte dans lequel il se déroule. La réussite d'un tel processus suppose que l'humain soit capable d'isoler un mouvement ou une séquence de postures en lui donnant un sens tout en intégrant les différences morphologiques, l'indépendance par rapport au point de vue et arrive même à se détacher du contexte.
La séquence généralement admise est celle correspondant à une suite de quatre phases [2, 4]:

1- Position de repos
2- La préparation (pre-stroke)
3- Le cœur (stroke)
4- La rétraction (post-stroke)

Cette séquence décrit la structure du geste. La difficulté majeure ici, consiste à déterminer les instants exacts où démarrent chacune des phases. Généralement, le problème de l'extraction du mouvement se confond avec le problème de reconnaissance. Une grande majorité de travaux concatènent la segmentation et la reconnaissance et les classifieurs sont utilisés autant pour reconnaître le geste que pour l'isoler du flot de mouvement continu [5, 6, 7]. Dans d'autres travaux, l'extraction effectue une recherche explicite du début et de la fin du mouvement. Kohol et Al. [8], réalisent une segmentation hiérarchique basée sur la rupture d'un modèle construit à partir des mesures de vitesse, d'accélération des parties du corps. Dans [9], une technique basée sur des filtres à particules permet de traquer les parties du haut du corps et d'estimer les mouvements de chacune d'elles.
En ce qui concerne le codage, la majorité des travaux considèrent le corps comme un système poly-articulé et lui applique les outils de robotique (variables articulaires, inerties, etc…). Cependant, pour certaines techniques s'appuyant sur le traitement d'image, le codage est sensiblement différent du fait de la non accessibilité aux données géométriques 3D. Dans [13] par exemple, le mouvement est codé comme un histogramme accumulant une transformé de Radon de la silhouette. Dans [9], les auteurs utilisent des indices cinématiques dérivés du flot optique, indices qu'ils classifient par la suite en considérant les 'sacs' de modes cinématiques.

2.3 Reconnaissance du geste

La dernière partie consiste à labéliser les segments de mouvements observés (dans le cas de la simultanéité de la segmentation et de la classification) ou extraits dans les autres. Cette labellisation est précédée par une phase d'apprentissage qui peut être supervisée ou non.
Dans une grande majorité de travaux, les chaines de Markov (HMM) sont utilisées. Cette dominance provient du fait que la structure et le principe même des HMM correspondent à la structure et à la dynamique des gestes.

Notre travail porte sur la reconnaissance en ligne d'un ensemble fini de 5 classes de gestes emblématiques. Dans ce qui suit, nous détaillons la technique que l'on a développée pour extraire les gestes présents dans un flot continu de mouvements. Par la suite, nous donnons quelques éléments concernant les classifieurs que nous avons au point pour reconnaître les gestes exécutés par des utilisateurs quelconques. Nous terminons par une description d'une application concernant le contrôle gestuel d'un robot mobile.

Codage et Segmentation des gestes

Dans notre travail, nous avons voulu un système qui soit indépendant de l'utilisateur. En d'autres termes, le système doit être robuste aux diverses variations citées avant (morphologie, tonicité, forme du geste). Pour ce faire, nous avons mis au point une technique de normalisation qui répond à ces besoins. En effet, pour construire nos descripteurs de geste, nous avons utilisé les vecteurs unitaires de chacune des parties en mouvement au lieu d'utiliser directement les variables articulaires (qui naturellement élimine la variable morphologique). Par

ailleurs, ce codage permet de normaliser la trajectoire par rapport au temps. En effet, un geste peut être vu comme une trajectoire atemporelle dans l'espace des configurations : le geste devient une courbe curviligne ou un parcours des positions dans l'espace des configurations (voir figure 2) indépendant du temps.

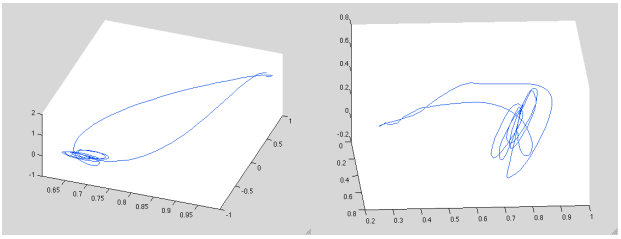


Figure 2: (a) Le geste BYE-BYE dans le volume (Z1, Z2, Z3) (b) le BYE-BYE dans le volume (X1, X2, X3).

Ceci nous permet de nous affranchir des éventuelles variations liées à la vitesse d'exécution ou la tonicité musculaire de l'exécutant. Malheureusement, ceci n'est pas suffisant. Comme on peut le constater sur la figure, pour les gestes cycliques, le nombre de boucles varie d'une personne à l'autre. Ce qui de fait enlèverait toute possibilité d'utiliser des techniques de normalisation telle que la DTW (Dynamic Time Warping) par exemple.

3.1 Codage des données brutes

Les gestes que nous considérons sont effectués par le bras droit. Aussi, nous utilisons les positions 3D des articulations (épaule, coude, poignet, et index). Ces dernières sont fournies indifféremment par un système de capture avec marqueurs (CODAMOTION ou PPT tracker) ou sans marqueurs (Kinect ou UpperBody Tracker [10]). Ces positions 3D sont d'abord transformées dans un repère lié au corps, notamment par rapport à la ligne joignant les deux épaules pour tenir compte de l'orientation relative du capteur par rapport à l'utilisateur. Par la suite, chaque segment est remplacé par son vecteur unitaire.

$$\overrightarrow{v_i} = \frac{\overrightarrow{A_i A_{i+1}}}{\|A_i A_{i+1}\|}.$$

Où, $\overrightarrow{v_i} = (v_{ix}, v_{iy}, v_{iz})$ est le vecteur normalisé entre les articulations (i) et (i+1). Nous avons alors 9 équations représentant une posture du bras droit. En plus de la normalisation d'échelle, nous avons aussi la possibilité d'éliminer le temps dans notre description et les 9 séries temporelles peuvent être exprimées comme des suites de postures P_i dans l'espace des configurations :

$$G = \left\{P_i, P_i = \left(v_{x,k=1,3}(i) \quad v_{y,k=1,3}(i) \quad v_{z,k=1,3}(i)\right)\right\}$$

Ainsi, il on peut normaliser les trajectoires par rapport au temps en ne considérant plus que les distances curvilignes dans l'espace des configurations indépendamment de la vitesse d'exécution des gestes (la tonicité de des utilisateurs).

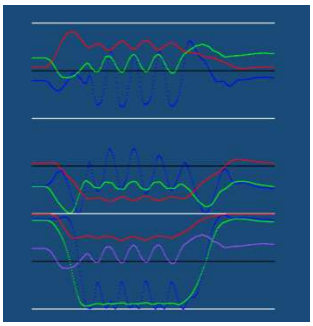


Figure 3: Les 9 composantes $v(t)_x$, $v(t)_y$, $v(t)_z$.

3.2 Segmentation

A partir des composantes normalisées, nous effectuons la segmentation du signal continu pour en extraire les parties correspondant à des gestes. La difficulté ici réside dans le fait qu'il est difficile de trouver un critère objectif pour déterminer le début et la fin d'une séquence candidate. Cette opération est complexe même pour un humain et elle reste incertaine [6]. Dans notre cas, nous effectuons la segmentation sur la base de propriétés cinématiques. En effet, le préparation du mouvement (le pre-stroke au sens de Kendon) consiste en un mouvement balistique qui emmène le(s) bras vers le cœur/corps du mouvement. Cette balistique consiste en une accélération puis une décélération à l'approche de la pose finale, puis une accélération et une décélération symétriques aux premières pour revenir à la position de repos (figure 3).

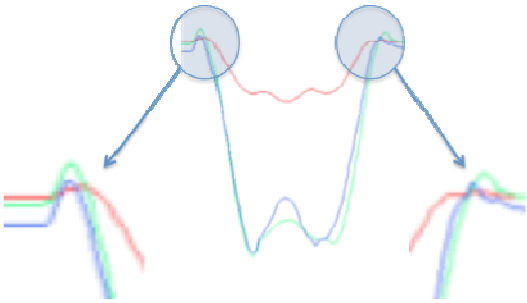


Figure 4: Le mouvement balistique dans le geste 'STOP'.

C'est cette forme caractéristique que nous utilisons pour avoir la première hypothèse sur le début et la fin du mouvement. Une seconde hypothèse concerne l'antiphase qui caractérise les mouvements : le bras part en sens inverse du mouvement à effectuer (zoom de la figure 2). A notre connaissance, cet indice n'a pas été utilisé auparavant car il est du même niveau que les fluctuations locales (mini mouvements du bras autour de la position de repos). Les bornes du mouvements sont obtenues à partir des extrema de la dérivée des signaux $v(t)_x$, $v(t)_y$, $v(t)_z$ confirmés par la présence du dépassement. La combinaison des deux indices nous permet un taux de

détection très élevé.

3.3 Codage du geste

A partir du signal segmenté, nous dérivons des descripteurs qui encodent les gestes. Cet encodage s'inspire en partie de la classification de Stokoe, pour qui le lieu du corps du geste (stroke) est une donnée caractéristique. En fait nous ne considérons pas le lieu géométrique de la posture au sens euclidien mais plutôt dans l'espace des configurations. En effet, le stroke correspond à un 'plateau' ou la posture est maintenue pendant un certain temps avant de repartir vers la position de repos. Ceci est valide aussi pour les gestes cycliques où les cycles s'effectuent autour d'une posture centrale.
La deuxième caractéristique que nous considérons est la '*cyclicité*'. Des gestes tels que ('VIENS', 'PASSE', 'BYE-BYE') ont une partie cyclique qu'il est facile de déterminer. En effet, l'utilisation d'un modèle autorégressif par exemple permettrait d'extraire cette caractéristique. Malheureusement, les cycles que nous avons observés ne sont pas des sinusoïdes pures (la personne ne reproduit pas exactement les mêmes parcours au sein d'un même geste). Pour tenir compte de cette variabilité, nous utilisons une distance qui mesure la différence entre un signal triangulaire et le mouvement effectué

$$P_{iso} = 1 - \prod_{i,j} \left(1 - \exp\left(-k_1 * d_{1ij}\right)\right) * \exp\left(-k_2 * d_{2ij}\right)$$

Où

$d_{1ij} = \left[0.5 * \left(m_i + m_{i+1}\right) - M_j\right]^2$ et $d_{2ij} = \left[\left(m_i - m_{i+1}\right)\right]^2$.

$k_{1,2}$ sont des constantes de normalisation.

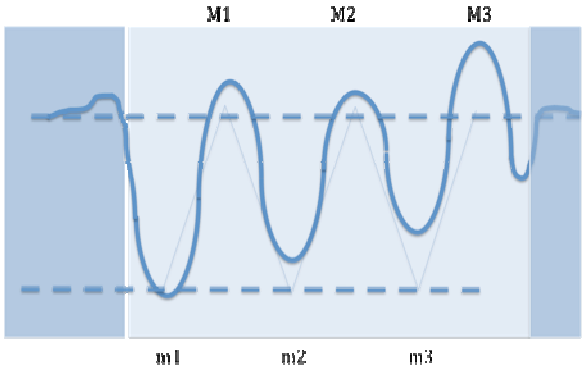


Figure 5: La pseudo-distance mesurant l'écart entre un geste cyclique théorique et un geste effectué.

La troisième et dernière caractéristique discriminante concerne l'amplitude des cycles. Cette dernière est simplement obtenue à partir des moyennes des différences entre extrema observés.
Ainsi, la série temporelle initiale est transformée en un vecteur de 27 composantes qui sont indépendantes de la dynamique du geste, de ses amplitudes ainsi que des variations locales.

Reconnaissance des gestes

La reconnaissance de gestes d'une manière générale est largement abordée par le biais Markovien (HMM Hiden Markov Model). Ces derniers sont parfaitement adaptés à la structure temporelle des gestes ainsi qu'à leurs formes. Dans cette contribution, nous avons utilisé une approche basée sur les SVM (Support Vector Machine). Ce choix a été dicté par notre problème : la classification des gestes s'apparente à un problème multi-classes où le classifieur doit fournir en sortie un score équivalent à une probabilité d'appartenance à une classe donnée. Plus spécifiquement, nous avons adopté l'implémentation giniSVM proposée dans [11]. Ce classifieur est un classifieur supervisé et à large marge. Son originalité est le fait qu'il combine une formulation quadratique de l'entropie et l'utilisation d'une distance quadratique pour le noyau, ce qui permet d'inclure une procédure de normalisation de la marge (au contraire des méthode classiques qui génèrent des probabilités conditionnelles biaisées et non normalisées). Les deux autres avantages du giniSVM sont, d'une part sa capacité à prendre en charge des données éparses (ce qui est le cas de nos données) et la rapidité des calculs pour trouver le vecteur support d'autre part.

Pour rappel, dans sa version bi-classes, la technique SVM est une technique d'apprentissage/classification qui consiste à trouver l'hyperplan optimal d'équation (w,b) qui subdivise un ensemble de données en deux sous-ensembles, ce qui reviendrait à minimiser la forme suivante :

$$\min_{w \in R^d} \|w\|^2$$

Sous la contrainte $y_i\left(< w, x_i > + b\right) \geq 1$. Dans le cas où on tolérerait une marge d'erreur (les sous-ensembles ne sont pas formellement linéairement séparables), la forme précédente devient :

$$\min_{w \in R^d} \|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

Avec la contrainte : $y_i\left(< w, x_i > + b\right) \geq 1 - \xi_i$. Où ξ_i est la variable ressort et C une constante de régularisation qui permet de gérer la marge d'erreur (la tolérance de retrouver des échantillons mal labélisés).
Dans le cas plus général de problèmes non séparables, il existe une technique le « *kernel trick* » qui permet de trouver des séparateurs grâce à la possibilité d'exprimer le problème dans des espaces de plus grandes dimensions à l'aide de transformations non linéaires : fonctions noyaux. Ce passage a plusieurs avantages, entre autres, il permet de conserver une forme quadratique, avec des métriques explicites et la possibilité d'appliquer des techniques d'optimisation classique.

Dans notre implémentation, nous avons utilisé une formulation proposée par Chakrabarty et Cauwenberghs :

le giniSVM. Cette technique transforme la formulation SVM classique de type 'grande marge' en un problème d'optimisation d'une fonction d'entropie. Cette dernière s'appuie sur une distance agnostique (au lieu de la distance euclidienne du SVM) pour fournir des distributions de probabilités conditionnelles normalisées et sans biais, qui peuvent être vues comme les vecteurs supports. Après apprentissage, ces derniers fournissent directement des probabilités d'appartenance aux classes.
Dans la pratique, nous avons implémenté le giniSVM avec une fonction noyau de type polynomial.
Avec cette implémentation, nous avons obtenu des résultats assez probants tant au niveau du temps de calcul (apprentissage et classification) qu'au niveau des taux d'erreurs de classification (voir section la 5.2 pour plus de détail).

Contrôle d'un robot

Comme nous l'avions annoncé précédemment, le système de reconnaissance de gestes est utilisé pour contrôler en ligne un robot à roues. Pour ce faire, nous avons développé un band d'essai comprenant :

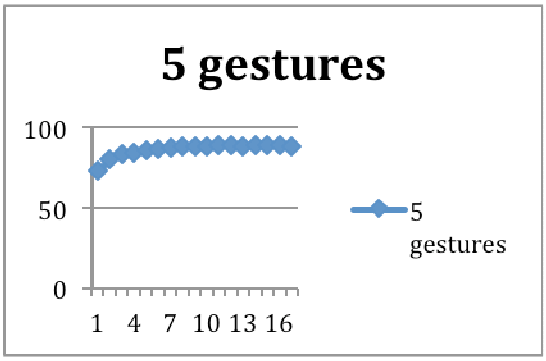
1- Robot mobile à roues contrôlable via une interface (ViRAT [12]). Le robot possède l'ensemble des fonctions de base lui assurant l'autonomie de naviguer sur la base d'un plan d'actions. Il est munit d'un LASER qui lui sert à éviter les obstacles, à construire un plan local (SLAM), etc.
2- Un serveur VRPN renvoyant la position d'un utilisateur comprenant une Kinect et un PC embarqué.
3- Un client connecté au serveur précédent qui effectue l'ensemble des calculs déportés aussi pour reconnaître les gestes et les envoyer au robot.

Figure 6 : Le robot Thumbler embarquant une Kinect.

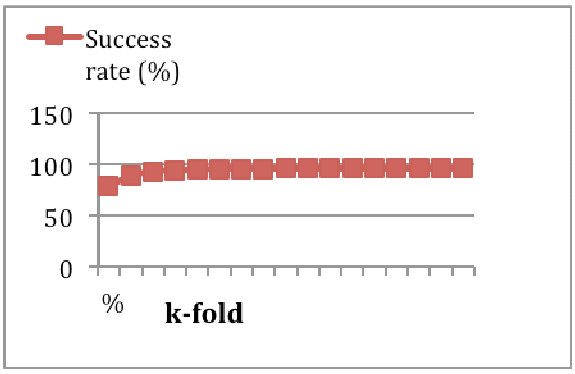
Figure 7 : Un utilisateur désignant une cible au robot.

1.1 Résultats et commentaires concernant la reconnaissance de gestes

Pour tester notre système de reconnaissance, nous avons utilisé 5 classes de gestes : le 'BYE-BYE' (classe 1), le 'STOP' (classe 2), le 'POINTAGE' (classe 3), le 'VIENS' (classe 4) et 'AVANCE' (classe5). Une population de 19 personnes nous a permis de recueillir, à partir d'un CODAMOTION, les données sur lesquels nous avons travaillé (10 répétitions*5 gestes *19 personnes). A cette population il était demandé simplement verbalement d'exécuter un geste donné : aucune instruction particulière n'a été donnée et nous voulions que les personnes exécutent les gestes aussi naturellement que possible.
Le choix des gestes correspond d'une part aux contrôles que les utilisateurs transmettront au robot. D'autre part, ces gestes permettent aussi d'avoir une mesure objective de la discrimination du module de reconnaissance. En effet, les trois gestes (1, 2 et 3) sont assez proches et constituent généralement des sources d'ambigüité.
Les résultats sont assez bons d'une manière générale et se situent autour des 92% (avec 5 gestes) et 95% (pour 4 gestes). Ce qui reste un taux tolérable pour des utilisations interactives.
La différence entre les deux séries viens de l'ambigüité entre deux gestes : le 'BYE-BYE' et le 'VIENS'. La matrice de confusion obtenue montre clairement une confusion entre eux car ils diffèrent simplement par l'orientation des plans où les cycles ont lieu (mouvements de la main). En théorie ces plans sont orthogonaux, mais selon les personnes, l'angle entre les deux plans n'est pas assez élevé et a induit en erreur notre classifieur.



Tab 1 : Taux de reconnaissance pour une librairie de 5 gestes en fonction de la base d'apprentissage



Tab 2 : Taux de reconnaissance pour une librairie de 4 gestes (le 'VIENS' en moins par rapport au tab précédent) en fonction de la base d'apprentissage

Une fois les tests hors-ligne réalisés, nous avons implémenté l'ensemble du module de reconnaissance pour supporter une Kinect qui a pour fonction de fournir le squelette de tout utilisateur qui se présente en face. Le module de segmentation en ligne a été testé avec un succès relatif (autour de 75%) : les données géométriques concernant squelette fournit par la Kinect sont assez bruitées (absence de quelques données et mauvaises estimations de la pose certainement en lien avec des artefacts que le capteur a du mal à appréhender). Pour ce qui est de la reconnaissance, nous avons des taux similaires aux taux obtenus hors-ligne ce qui nous donne une indication sur la robustesse de notre approche.

5.2 Contrôles du robot

Nous avons réalisé des tests en laboratoire pour valider l'ensemble de la chaine. L'objectif attendu n'était pas de tester l'utilisabilité du système mais, de tester sa cohérence et la faisabilité du contrôle notamment dans un contexte d'interaction, où la boucle de contrôle doit être aussi rapide que possible. A ce propos, nous avons un retard maximum d'une seconde, qui s'explique par le retard induit par le capteur (production du squelette et transmission à travers le réseau sans-fil) ainsi que par le système de traitement.

5.3 Synthèse de gestes

Notre travail n'a pas traité directement de la synthèse de gestes plausibles. Notre objectif premier était de fournir un système capable de contrôler un robot à l'aide de gestes naturels. Cependant, les résultats obtenus pour la reconnaissance minimaliste à l'aide d'un classifieur de type SVM peuvent être potentiellement intéressants dans une application de synthèse. D'une part nous proposons un codage invariant par rapport à la morphologie, la tonicité, et du point de vue du receveur : nous obtenons une formulation compacte et adimensionnelle du geste. Cette formulation peut être interprétée de deux façons :
1) **mécaniste** avec des composantes qui décrivent la relation du bras avec son environnement et le reste du corps : directions par rapport à la verticale et distances par rapport au corps.
2) **visuelle** avec des composantes qui approchent au mieux la forme apparente de la partie du bras vue par le receveur : un avant bras à la verticale sera vu dans sa totalité alors qu'un bras tendu (dans un pointing par exemple) sera vu partiellement.

Par ailleurs, les indices que nous utilisons pour représenter les gestes sont directement liés à la notion de forme ou morphologie du geste. Nous ne pouvons affirmer ici que ces indices peuvent être pris comme un système de codage réversible, mais il facile d'imaginer qu'il est possible de profiler les trajectoires normalisées en modifiant la tonicité en effectuant une DTW (dynamic time warping) par exemple, en modifiant la taille de l'exécutant en modifiant un facteur d'échelle. Ce qui par rapport aux techniques actuelles de synthèse et d'apprentissage du geste (stochastiques dans [16] par exemple) permet de plus grandes latitudes dans le formatage du geste.

Conclusion et travaux futurs

Dans ce papier, nous avons présenté les résultats préliminaires de l'intégration d'un module de reconnaissance de gestes en ligne appliqué au contrôle d'un robot mobile. Les résultats préliminaires sont encourageants : les taux de réussite dans la segmentation et dans la reconnaissance sont assez élevés par rapport à la littérature. En effet, l'apprentissage du système a été effectué sur un nombre réduit de sujets ce qui n'a pas empêcher des utilisateurs ne faisant pas partie de la base d'apprentissage d'être reconnus. Il est clair que notre système n'en est qu'à ses débuts. Pour l'instant nous avons implémenté uniquement un jeu de 5 gestes, correspondant aux fonctions de bases de contrôle du mouvement du robot. De nombreux autres gestes peuvent être facilement appris pour d'autres scénarii. Nous avons vu qu'il y a une ambigüité entre deux gestes. Nous pensons que cette dernière peut être levée en affinant plus la transformation du bras par rapport au corps : pour l'instant, nous considérons l'orientation relative du bras

par rapport à la ligne des épaules et non au corps entier. Un autre souci concerne la segmentation en ligne : le taux actuel n'est pas acceptable et un filtrage des données brutes s'impose. Par ailleurs, la limitation du nombre de gestes peut être aisément dépassée en intégrant de nouveaux gestes dans la base d'apprentissage. A cet effet, nous prévoyons un schéma d'apprentissage non supervisé afin de pouvoir intégrer de nouveaux gestes d'une façon automatique et sans interventions. Une autre amélioration consiste à intégrer la voix pour d'une part lever les ambigüités concernant la reconnaissance de gestes, et de se placer dans une optique multimodale qui prend en charge les gesticulations (combinaison de gestes et de paroles) surtout pour les séquences de pointage (désignation vocale d'un objet pour lever l'ambigüité dans un environnement encombré).

Enfin, nous avons abordé l'aspect synthèse de gestes plutôt dans sa dimension perspective. Dans ce travail, nous avons montré qu'un geste pouvait être transformé dans une forme adimensionnelle. Cette forme permet, selon toute vraisemblance, de supporter aussi bien la reconnaissance que la synthèse avec de la modulation 'morpho-kinetics'. Un travail sur le dernier aspect permettra certainement d'arriver à la création d'animations plausibles dans le cadre des agents virtuels.

Remerciements

Les auteurs remercient les organisateurs du workshop pour les suggestions importantes qu'ils ont formulé à l'encontre de la version préliminaire de ce travail.

Bibliographie

[1] A. Kendon, An Agenda for Gesture Studie, *Semiotic Review of Books*, Volume 7 (3), pp 8-12, 1996.

[2] A. Corradini et P.R Cohen, Speech_gesture Interface for Handfree Painting on a Virtual Paper suing Partial Neural Networks as Gesture Recognizer, in *Proceedings IJCNN'02, HI, 2002 2293-2298*.

[3] D. McNeill, Hand and Mind. What Gestures Reveal about Thought. Chicago University Press. 1992.

[4] M. Karam and M. C. Schraefel. A taxonomy of gestures in human computer interaction. *Technical report, Electronics and Computer Science, University of Southampton*, 2005. vii, 69, 70, 71.

[5] D. Kim, J. Song, D. Kim. Simultaneous gesture segmentation and recognition based on forward spotting accumulative HMMs. In "*Pattern Recognition*". Vol 40 (2007) pp 3012–3026.

[6] M. Sigalas, H. Baltzakis, and P. Trahanias. Gesture recognition based on arm tracking for human-robot interaction. *In Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, October 2010.

[7] R. Billon, A. Nédélec, J. Tisseau. Gesture Recognition in Flow in the Context of Virtual Theater. In *Proceedings of IVA'2008*. pp.470~471 .

[8] K. Kahol, Kanav, P. Tripathi, and S Panchanathan. Gesture segmentation in complex motion sequences. *Proceedings 2003 International Conference on Image Processing*.

[9] S. Ali, M. Shah. Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning, *IEEE Transactions on* Pattern Analysis and Machine Intelligence, Vol 32, issue 2, pp 288 – 303. 2010.

[10] R. Chellali, I. Renna, C. Achard, Emblematic Gestures recognition. A paraître dans proceedings EASD ASME conference, July 2012.

[11] S. Chakrabartty et Gert Cauwenberghs. Gini Support Vector Machine: Quadratic Entropy Based Robust Multi-Class Probability Regression, *Journal of Machine Learning Research*, 8 (2007) 813-839.

[12] W. C Stokoe, D. C. Casterline, C. G. Croneberg (1965) *A Dictionary of American Sign Language on Linguistic Principles*, 2nd edition.

[13] E. Suma, C.W. Sinclair, J. Babbs, and R. Souvenir, A Sketch-Based Approach for Detecting Common Human Actions, *International Symposium on Visual Computing 2008*, pp. 418-427.

[14] C. Pelachaud, Studies on Gesture Expressivity for a Virtual Agent, Speech Communication, *Special issue in honor of Björn Granstrom and Rolf Carlson, 51, 2009*, 630-639

[15] Grafton and A. Hamilton. Evidence for a distributed hierarchy of action représentation in the brain. *Human Motor Sciences*, pages 590–616, 2007.

[16] Calinon, S.; Sauser, E.L.; Billard, A.G.; Caldwell, D.G., Evaluation of a probabilistic approach to learn and reproduce gestures by imitation, IEEE ICRA 2010, pp2671 - 2676